University of Technology Sydney

Doctoral Thesis

Recognising and Describing Human Activities in a Still Image

Author:
Zheng Zhou

Supervisor:
Xiangjian He

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

School of Computing and Communications
Faculty of Engineering and Information Technology

September 2017

Declaration of Authorship

This thesis is the result of a research candidate conducted jointly with another University as part of a collaborative Doctoral degree. I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signed: _____

Date: _____

Abstract

Understanding human activities in a still image is an essential research branch of artificial intelligence. For a computer system, the ability of understanding activities in an image is composed of not only the ability of recognising the activities in the image but also the ability of describing the recognised activities. In the age of big data, activity recognition and description generation for image have received increasing research attention, since they are of great importance in image-based information retrieval, automated image collection and collation, human-computer interaction and automated security surveillance. This thesis conducts research on recognising and describing activities in a still image and achieves several innovative achievements as follows.

(1) A framework for recognising human activities based on analysing the interactions among people is proposed. The interactions among people provide useful context for activity recognition but have not been fully taken advantage of by the existing approaches for both individual and group activity recognition. The framework is constructed based on analysing the mechanism that human brains analyse the interactions, and composed of four key sub-tasks, including Human Detection and Segmentation, Feature Extraction, Interaction Analysis and Activity Recognition.

(2) An approach for recognising individual activities based on human-interaction analysis is developed. This approach uses an innovative single-level model, called the Non-hierarchical Interaction Analysis Model (NIAM), to analyse the interactions between individuals. The NIAM does not contain a level representing groups and a group discovery process, in order to avoid the errors occurred in and computation consumed for group discovery. Several innovative algorithms are proposed and compose the body of the recognition approach, including a Fusion Restricted Boltzmann Machine for fusing features of different dimensional scales, a Focal Subspace Measurement for calculating the interdependencies between people and a Global-Local Cue Integration Method for selecting and integrating the cues extracted from different people.

(3) An approach for recognising group activities based on human-interaction analysis is developed. This approach uses a new multiple-level generative model, called Mixed Group Activity Model (MGAM). Compared with the popular discriminative multiple-level models, the MGAM performs better in comprehensively analysing the information of multiple levels of activities and modeling the interactions among multiple individuals or groups. To connect the MGAM with the raw features in an image, a Body-Part-Angle (BPA) descriptor is proposed. The BPA descriptor is friendly to a generative model that the generation distribution between the model and the raw features can be easily defined and learned.

(4) A description generator for describing the human-object interaction activities in images with natural language is proposed. Compared with the sentences given by the traditional retrieval-based approaches, the sentences given by this generator are closer to what is really happening in an image. The generator is implemented based on a deep understanding framework with a 3D spatial layout analysis and a syntactic-tree-based language model.

Acknowledgements

Foremost, I wish to appreciate my principle supervisor Professor Xiangjian He for his patience, motivation, constant encouragement and help during my candidature. His guidance helped me in all the time of research. Without his support, I cannot finish my PhD study.

I would also like to thank my colleagues and the staff in the Faculty of Engineering and Information Technology (FEIT): Wenjing Jia, Qiang Wu, Min Xu, Haiying Xia, Fan Dong, Shuai Jiang, Khaled Aldebei, etc., for their invaluable help and support. I have the honour of studying and working with them in the past two years which is valuably stamped in my life.

Last but not least, I would like to thank my family for their unconditional support throughout my life. I met my wife Jianghan Ke in the period of my PhD study. She was the biggest treasure that I gained during my PhD time and will be the biggest treasure that I ever have during my whole life.

Contents

List of Figures

List of Tables